A Survey of Machine Learning Methods for Big Data

Zoila Ruiz[1], Jaime Salvador[1], and Jose Garcia-Rodriguez[2(✉)]

[1] Universidad Central Del Ecuador, Ciudadela Universitaria, Quito, Ecuador
{zruiz,jsalvador}@uce.edu.ec
[2] Universidad de Alicante, Ap. 99, 03080 Alicante, Spain
jgarcia@dtic.ua.es

Abstract. Nowadays there are studies in different fields aimed to extract relevant information on trends, challenges and opportunities; all these studies have something in common: they work with large volumes of data. This work analyzes different studies carried out on the use of Machine Learning (ML) for processing large volumes of data (Big Data). Most of these datasets, are complex and come from various sources with structured or unstructured data. For this reason, it is necessary to find mechanisms that allow classification and, in a certain way, organize them to facilitate to the users the extraction of the required information. The processing of these data requires the use of classification techniques that will also be reviewed.

Keywords: Big Data · Machine learning · Classification · Clustering

1 Introduction

In recent years there has been an accelerated growth in the volume of information available on the network. Likewise, several alternatives have appeared for processing these large volumes of data (Big Data) and their storage [18]. There are many studies oriented to the processing of Big Data and extraction of relevant information that allows to generate knowledge [17]. The different techniques of Machine Learning allow to achieve this purpose, therefore, several studies have been devoted to replace the statistical analysis by ML techniques. For this reason, it's necessary to study the main data characteristics, such as: heterogeneity, autonomy, complexity and evolution [40].

Nowadays, presenting alternatives for processing large volumes of data is undoubtedly the main goal of many works or projects [32]. The ability to process information with savings in computational costs "maximizing the relevance and reliability of the information obtained" is a real challenge [2]. Machine Learning techniques are a good alternative to solve problems related to Big Data. In this document we review some algorithms used in different works, combination of techniques or proposals of hybrid techniques used to process large volumes of data.

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2017, Part II, LNCS 10338, pp. 259–267, 2017.
DOI: 10.1007/978-3-319-59773-7_27

This document is organized as follows. Section 2 introduces the Big Data concept summarizing its characteristics; later, some classification techniques and algorithms are described. Section 3 presents a discussion about the criteria for choosing a classification technique. Finally, Sect. 4 presents some conclusions and opportunities for future work.

2 Big Data

This section describes the most common datasets used for verifying certain algorithms. Some of the most relevant classification algorithms are reviewed below.

Big Data is present in all areas and sectors worldwide. However, it's complexity exceeds the processing power of traditional tools. Because of this, high-performance computing platforms are required to exploit the full power of Big Data [36]. These requirements have undoubtedly become a real challenge. Many studies focus on the search of methodologies that allow lowering computational costs with an increase in the relevance of extracted information. The need to extract useful knowledge has required researchers to apply different machine learning techniques, to compare the results obtained and to analyze them according to the characteristics of the large data volumes (volume, velocity,veracity and variety, the 4V's) [26].

2.1 Datasets

With the growth of data size, it is essential to consider techniques to find complex relationships between samples and models always considering the evolution of data over time [39]. In this way, we can build systems whose design allows unstructured data to be linked through relationships. This will allow us to obtain valid patterns through which trends can be predicted or a phenomenon can be better understood.

Table 1 briefly describes some types of popular datasets used to validate different methods commonly used in processing large volumes of data.

Table 1. Features present in popular datasets

	Features		
Dataset	Velocity	Volume	Variety
Repository of the ML databases [16]		X	X
Social computing [26]	X	X	X
Synthetic interval datasets [9]	X	X	
Sociodemographic data [24]		X	X
Real database [12]	X	X	X

As we observed in the previous table, there are datasets that can be used for the verification, validation, comparison and previous training of the algorithms

to process the data. Many of these algorithms require training to properly process the data. Each dataset has features that allow you to choose the ones that best fit the actual data [30].

2.2 Classification Techniques

In this section we introduce the most relevant algorithms for classification and its relationship with Big Data platforms. First, a classification of machine learning techniques is presented, later we describe some classification algorithms.

The classification algorithms are divided into:

- *Supervised.* The main task is to determine to which class each new data belongs. This is achieved based on a training-classification scheme using previously established sample sets. These techniques can only be used if the number of classes is known a priori. Examples of these algorithms are Neighborhood Based, Decision Trees (DT) and Support Vector Machines (SVM).
- *Unsupervised.* They are used when training sets are not available. Therefore, they use grouping algorithms to construct groups, so that the data belonging to the group has a high level of similarity. Among the most used algorithms we can find K-Means, Sequential Grouping, ISODATA or Adaptive method.

The main problem found in different studies oriented to the processing of large volumes of data resides in the selection of suitable techniques for variable selection and classification. The technique chosen depends on the type of information analyzed, this allows to obtain higher quality information, reduces the computational cost and improves processing times. Some of the most used criteria are: the dimensionality of the data, the relevant features [14] and the veracity of the information obtained. With these considerations we can select the most appropriate Machine Learning techniques that allow us to optimize the results obtained.

2.3 Classification Algorithms

The following are some of the most relevant classification algorithms:

- *K-Means:* Simple and efficient, it needs only one initial parameter (k) and its results depend on the initial selection of the clusters centroids [1].
- *K-Medoids:* It is considered a K-Means variation. Its goal is to determine the best representative of the center of each cluster (medoide) [29].
- *Support Vector Machine (SVM):* Given a training set with a labeled class (through training), SVM can build a model that predicts the class of a new sample [31].
- *k-Nearest Neighbor (KNN):* It is simple and local. You need to specify an appropriate metric to measure proximity. It is noise and dimensionality sensitive [7].

- *Expectation-Maximization (EM):* It provides a maximum likelihood iterative solution. It converges to a local maximum and is sensitive to the choice of the initial values [20].
- *Self Organized Map (SOM):* It groups data from the input set according to different criteria from a training process. An intuitive description of the similarity between data can be observed through a map [22].
- *DBSCAN:* Automatically determines the number of clusters. Points with low density are classified as noise and are omitted, so there is no complete clustering [6].
- *Decision Tree (DT):* It is recursevely constructed following a hierarchical descending strategy [11].

In several works we find "as the used methodology" a variation of these Machine Learning algorithms. These variations allow, in a certain way, to eliminate or minimize the limitations present in each one. Sometimes they depend directly on the data set used in the experimentation stage or on the initialization parameters of each algorithm, among others. In other cases they propose a technique [42] or hybrid strategy for processing large volumes of data.

Table 2 summarizes different studies that propose the combination of algorithms, metrics or pre-processing of data using another algorithm. This contributes to improving the processing of information (minimizing computational costs and maximizing the relevance of extracted information).

Table 2. Combining algorithms

Algorithm	EM	Fuzzy	PSO[a]	Bisection	GA[b]	KNN
K-Means [10]	X	X	X	X	X	X
K-Medoids [35]		X	X		X	X
SVM [44]	X	X	X	X	X	X
KNN [13]	X	X	X		X	
SOM [21]	X	X	X		X	X
DBSCAN [15]		X	X		X	X
DT [37]	X	X	X		X	X

[a]Particle Swarm Optimization
[b] Genetic Algorithm

Table 3 shows some hybrid techniques or strategies. Different methods are used for similarity calculation. They combine different ML techniques to create more efficient classification methods.

In the reviewed papers, a distinction was found between implementing hybrid techniques or strategies. The first case is based on introducing in the algorithm some technique different from the one usually used for the internal calculation of some parameter. In the second case, a certain limitation of a technique is strengthened in the pre-processing stage. For example, if it is not suitable for

Table 3. Techniques and hybrid strategies

Technique	Hybrid techniques	Hybrid strategy
Hybrid Bisect K-Means [27]	X	X
HOPACH [a] [38]	X	
DHG [b][12]	X	X
K-mean and KHM [c] [19]	X	
K-Means - GA [3]	X	
HcGA [d] [5]		X
HFS [e] [43]		X
MAM - SOM		X
K-ICA [28]	X	
GKA [f][33]	X	
NKMC [g] [8]	X	
HSRS [h] [34]		X
HC-HOSVD [i] [23]	X	

[a]Hierarchical Ordered Partitioning And Collapsing Hybrid
[b]Density-based Hierarchical Gaussian
[c]K-Harmonic Mean
[d]Hybrid Cellular Genetic Algorithms
[e]Hybrid feature selection scheme
[f]Genetic K-means Algorithm
[g]Naive multi-view K-means
[h]Hybrid sequential-ranked searches
[i]Hybrid clustering via Higher-order singular value decomposition

processing large volumes of data, the data is first partitioned with an appropriate technique and then the resulting technique is applied to each resulting partition.

Among the most outstanding studies we can mention the one of Mishra and Raghavan [25], comparison of optimization algorithms, Al-Sultan and Khan [4], about algorithms like K-means, SA[1], TS[2] and GA, Xiaowei and Ester [41] among others.

From these works three main conclusions can be drawn:

- No method outperforms the other methods in regards to performance when working with one-dimensional or multi-dimensional data.
- Solutions found by TS, GA and SA outperform K-Means, but this is much faster. GA is the fastest finding the best solution, while SA is the fastest in converging.
- The problem with these algorithms is that they do not work properly for large volumes of data, only K-means and Kohonen Maps have been successfully applied to large datasets.

[1] Simulated Annealing.
[2] Tabu search.

3 Discussion

In this section we analyze the usual criteria taken into account when choosing a Machine Learning algorithm to efficiently process information. This allows a greater comprehensibility and interest of the extracted data.

The proposal of new perspectives in the classification methods allows to open new lines of investigation. Considering different metrics to establish similarity between groups or combine techniques for parameter adjustment, allow to improve the results obtained by directly applying a technique. The considerations that are analyzed prior to choosing a classification technique are:

- Data to be processed
- Limitations and parameters to each algorithm

3.1 Data to be Processed

The high dimensionaly of data features is one of the problems that must be considered when working with large volumes of data. The reduction of the data dimensionality is considered, for which the feature selection is very important; to take this into account, algorithms of feature selection and extraction are usually applied. Another valid consideration is the structure and specific features of the data.

3.2 Algorithms Limitations and Accurate Parameterization

Some algorithms are more suitable to process large volumes of data but are not necessarily faster to find the best solution or the least costly; however, by analyzing different algorithms we can obtain important considerations and propose improvements that exceed the limitations of each algorithm. For example, replacing the sequential search of the winning unit for a faster and more efficient search (MAM-SOM).

With this previous analysis, we can pre-select the most adequate techniques for processing our data. However, depending on the data characteristics and the planned objectives, in most cases it requires a combination of techniques, preprocessing of the data, modification of the inner calculations of the algorithms, or hybrid strategies to achieve optimal results.

To verify its behavior, performance and favorable parameters for optimal performance, it is necessary to evaluate each algorithm and compare the results through experimentation. The best way is to use different types of datasets, considering that there are algorithms that work better with categorical data and others do with quantitative data, but very few manage data that have both characteristics simultaneously.

We need to evaluate the validity, stability and scalability in the results obtained in each algorithm.

– Validity: Determining the precision of an algorithm for data clustering.
– Stability: The variation of results obtained in different executions must be similar to each other.
– Scalability: The capacity of clustering big volumes of data in an efficiently way.

4 Conclusions

Machine Learning algorithms have a series of advantages and disadvantages that are reflected in execution times, computational requirements, convergence capacity, complexity levels, their implementation or parameter adjustment among others. Therefore, in many studies it has been decided to combine algorithms to solve problems when processing large volumes of data, depending on their characteristics and goals. It is possible to take advantage of the characteristics of two or more techniques, at the same time, to provide versatile tools in the processing of Big Data.

Despite the existence of a large number of ML techniques, most have some limitations. Problems such as overlap between groups, presence of noise or irregular structures are usually treated using hybrid techniques or strategies. The replacement of an internal calculation by another ML technique allows to overcome the limitations of the algorithms.

As future work, it is intended to test each technique and combination presented with public datasets on various platforms. To test the scalability, computational cost and response time of the different techniques. We should consider parallelization as an alternative to improve aspects such as: response time, computational capacity required by some algorithms or the ability to process large volumes of data.

Acknowledgements. This work has been funded by the Spanish Government TIN2016-76515-R grant for the COMBAHO project, supported with Feder funds.

References

1. Agrawal, A.: Global K-means (GKM) clustering algorithm: a survey. Int. J. Comput. Appl. **79**(2), 20–24 (2013)
2. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K.: Efficient machine learning for big data: a review. Big Data Res. **2**(3), 87–93 (2015)
3. Al Malki, A., Rizk, M.M., El-Shorbagy, M.A., Mousa, A.A., Malki, A.A., Rizk, M.M., Mousa, A.A., Mousa, A.A.: Hybrid genetic algorithm with K-means for clustering problems. Open J. Optim. **5**(02), 71 (2016)
4. Al-Sultana, K.S., Khan, M.M.: Computational experience on four algorithms for the hard clustering problem. Pattern Recogn. Lett. **17**(3), 295–308 (1996)
5. Arellano-Verdejo, J., Alba, E., Godoy-Calderon, S.: Efficiently finding the optimum number of clusters in a dataset with a new hybrid differential evolution algorithm: DELA. Soft. Comput. **20**(3), 895–905 (2016)

6. Backlund, H., Hedblom, A., Neijman, N.: A density-based spatial clustering of application with noise. Data Mining TNM033, pp. 11–30 (2011)
7. Bobadilla, J., Ortega, F., Hernando, A., de Rivera, G.G.: A similarity metric designed to speed up, using hardware, the recommender systems k-nearest neighbors algorithm. Knowl.-Based Syst. **51**, 27–34 (2013)
8. Cai, X., Nie, F., Huang, H.: Multi-view K-means clustering on big data. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 2598–2604 (2013)
9. De Carvalho, F.A.T.: Fuzzy c-means clustering methods for symbolic interval data. Pattern Recogn. Lett. **28**(4), 423–437 (2007)
10. Cui, X., Potok, T.E.: Document clustering analysis based on hybrid PSO + K-means algorithm. J. Comput. Sci. **27**(special issue), 33 (2005)
11. Dai, W., Ji, W.: A MapReduce implementation of C4. 5 decision tree algorithm. Int. J. Database Theory Appl. **7**(1), 49–60 (2014)
12. Pascual, D., Pla, F., Sánchez, J.S.: A density-based hierarchical clustering algorithm for highly overlapped distributions with noisy points. In: CCIA, vol. 220, pp. 183–192 (2010)
13. Derrac, J., Chiclana, F., García, S., Herrera, F.: Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets. Inf. Sci. **329**, 144–163 (2016)
14. Fan, W., Bifet, A.: Mining big data : current status, and forecast to the future. ACM SIGKDD Explor. Newsl. **14**(2), 1–5 (2013)
15. Feng, X., Wang, Z., Yin, G., Wang, Y.: PSO-based DBSCAN with obstacle constraints. J. Theor. Appl. Inf. Technol. **46**(1), 377–383 (2012)
16. Hatamlou, A.: Black hole: a new heuristic optimization approach for data clustering. Inf. Sci. **222**, 175–184 (2013)
17. Ho, R.: Big data machine learning: patterns for predictive analytics. DZone Refcardz **158**, 1–6 (2012)
18. Jadhav, D.K.: Big data: the new challenges in data mining. Int. J. Innov. Res. Comput. Sci. Technol. **1**(2), 39–42 (2013)
19. Jain, R.: A hybrid clustering algorithm for data mining, pp. 387–393 (2012). arXiv preprint arXiv:1205.5353
20. Jiang, M., Ding, Y., Goertzel, B., Huang, Z., Zhou, C., Chao, F.: Improving machine vision via incorporating expectation-maximization into deep spatio-temporal learning. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1804–1811 (2014)
21. Jin, H., Shum, W.-H., Leung, K.-S., Wong, M.-L.: Expanding self-organizing map for data visualization and cluster analysis. Inf. Sci. **163**(1–3), 157–173 (2004)
22. Kohonen, T.: Essentials of the self-organizing map. Neural Netw. **37**, 52–65 (2013)
23. Liu, X., Lathauwer, L., Janssens, F., Moor, B.: Hybrid clustering of multiple information sources via HOSVD. In: Zhang, L., Lu, B.-L., Kwok, J. (eds.) ISNN 2010. LNCS, vol. 6064, pp. 337–345. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13318-3_42
24. Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., Phung, D., Venkatesh, S., Allender, S.: Is demography destiny? application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. PLoS ONE **10**(5), e0125602 (2015)
25. Mishra, S.K., Raghavan, V.V.: An empirical study of the performance of heuristic methods for clustering. In: Pattern Recognition in Practice IV - Multiple Paradigms, Comparative Studies and Hybrid Systems, pp. 425–436. Elsevier BV (1994)
26. Mujeeb, S., Naidu, L.K.: A relative study on big data applications and techniques. Int. J. Eng. Innov. Technol. (IJEIT) **4**(10), 133–138 (2015)

27. Murugesan, K., Jun, Z.: Hybrid bisect K-means clustering algorithm. In: International Conference on Business Computing and Global Informatization (BCGIN), pp. 216–219. IEEE (2011)
28. Niknam, T., Fard, E.T., Pourjafarian, N., Rousta, A.: An efficient hybrid algorithm based on modified imperialist competitive algorithm and k-means for data clustering. Eng. Appl. Artif. Intell. **24**(2), 306–317 (2011)
29. Park, H.-S., Jun, C.-H.: A simple and fast algorithm for k-medoids clustering. Expert Syst. Appl. **36**(2), 3336–3341 (2009)
30. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data. ACM SIGKDD Explor. Newsl. **6**(1), 90–105 (2004)
31. Qi, Z., Tian, Y., Shi, Y.: Robust twin support vector machine for pattern classification. Pattern Recogn. **46**(1), 305–316 (2013)
32. Rebentrost, P., Mohseni, M., Lloyd, S.: Quantum support vector machine for big data classification. Phys. Rev. Lett. **113**(3), 1–5 (2014)
33. Roy, D.K., Sharma, L.K.: Genetic k-Means clustering algorithm for mixed numeric and categorical data sets. Int. J. Artif. Intell. Appl. **1**, 23–28 (2010)
34. Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S., García-Torres, M.: Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches. Expert Syst. Appl. **39**(12), 11094–11102 (2012)
35. Sheng, W., Liu, X.: A genetic k-medoids clustering algorithm. J. Heuristics **12**(6), 447–466 (2006)
36. Shim, K.: MapReduce algorithms for big data analysis. In: Madaan, A., Kikuchi, S., Bhalla, S. (eds.) DNIS 2013. LNCS, vol. 7813, pp. 44–48. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37134-9_3
37. Tsai, M.-C., Chen, K.-H., Su, C.-T., Lin, H.-C.: An Application of PSO algorithm and decision tree for medical problem. In: 2nd Internatonal Conference on Intelligent Computational System, pp. 124–126 (2012)
38. van der Laan, M.J., Pollard, K.S.: A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. J. Stat. Plann. Infer. **117**, 275–303 (2003)
39. Venkatesh, H., Perur, S.D., Jalihal, N.: A study on use of big data in cloud computing environment. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **6**(3), 2076–2078 (2015)
40. Wu, X., Zhu, X., Wu, G.-Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
41. Xu, X., Ester, M., Kriegel, H.-P., Sander, J.: A distribution-based clustering algorithm for mining in large spatial databases. In: 14th International Conference on Data Engineering (ICDE 1998) (1998)
42. Yang, F., Sun, T., Zhang, C.: An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization. Expert Syst. Appl. **36**(6), 9847–9852 (2009)
43. Yang, Y., Liao, Y., Meng, G., Lee, J.: A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. Expert Syst. Appl. **38**(9), 11311–11320 (2011)
44. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. **2**, 2126–2136 (2006)